J Mol Evol (2010) 70:605–612
DOI 10.1007/s00239-010-9356-9

Empirical Analysis of the Most Relevant Parameters of Codon Substitution Models

Stefan Zoller · Adrian Schneider

Received: 27 January 2010 / Accepted: 17 May 2010 / Published online: 5 June 2010
© Springer Science+Business Media, LLC 2010

Abstract Traditionally, codon models of evolution have been parametric, meaning that the 61 × 61 substitution rate matrix was derived from only a handful of parameters, typically the equilibrium frequencies, the ratio of nonsynonymous to synonymous substitution rates and the ratio between transition and transversion rates. These parameters are reasonable choices and are based on observations of what aspects of evolution often vary in coding DNA. However, the choices are relatively arbitrary and no systematic empirical search has ever been performed to identify the best parameters for a codon model. Even for the empirical or semi-empirical models that have been presented recently, only the average substitution rates have been estimated from databases of real coding DNA, but the parameters used were essentially the same as before. In this study we attempted to investigate empirically what the most relevant parameters for a codon model are. By performing a principal component analysis (PCA) on 3666 substitution rate matrices estimated from single gene families, the sets of the most co-varying substitution rates were determined. Interestingly, the two most significant principal components (PCs) describe clearly identifiable parameters: the first PC separates synonymous and nonsynonymous substitutions while the second PC distinguishes between substitutions where only one nucleotide changes and substitutions with two or three nucleotide changes. For the third and subsequent PCs no simple descriptions could be found.

Keywords Codon models · Markov models · Coding sequence evolution · Codon substitutions

S. Zoller · A. Schneider (✉)
Computer Science Department, ETH Zurich, Zürich, Switzerland
e-mail: schneadr@gmail.com

S. Zoller
e-mail: zollers@student.ethz.ch

Present Address:
A. Schneider
Institute of Evolutionary Biology, University of Edinburgh, West Mains Rd, Edinburgh EH9 3JT, UK

Introduction

The evolutionary changes in proteins can be modeled on amino acid or on DNA level. Normally, a rate matrix defining a Markov process is used to generate the substitution probabilities between any two residue states. When the modeling is performed on DNA, there is a further choice of either treating every nucleotide separately (4 states) or by assigning probabilities to the substitutions of nucleotide triplets, the codons (64 states, or 61 if the stop codons are ignored).

The first codon models of evolution were presented in 1994, simultaneously by Goldman and Yang as well as by Muse and Gaut. The two models are stated with different terms and parameters, but in the effect they are very similar. The former (Goldman and Yang 1994), however, became more widely used, probably due to its implementation in the *paml* software package (Yang 1997). Their 61 × 61 substitution rate matrix is derived from the codon frequencies π, the selection coefficient ω (the ratio of nonsynonymous to synonymous substitution rates), the ratio κ (relative rates of transitions and transversions) as well as a parameter V modeling the influence of the physico-chemical distances between different amino acids. However, the model is normally used in a simplified

version without the V parameter, allowing for a more direct interpretation of the ω value (Yang 1998).

These models are often called parametric models since all substitution rates are derived from parameters. This has the advantage that the parameter values can be estimated directly from the data set on which the analysis is performed and thus, the model can incorporate the particularities of the sequences under investigation. The parametric models have been widely used and also improved over the following years, for example by allowing ω to vary among sites (Nielsen and Yang 1998) or among lineages (Yang 1998) or both (Yang and Nielsen 2002). However, they still did not cover all known aspects of coding sequence evolution. In particular, the varying similarities among amino acids are often ignored and the rates of substitutions involving more than one nucleotide change are generally assumed to be zero.

In 2005, the first empirical codon model has been presented (Schneider et al. 2005), for which all substitution rates have been estimated from a large data set of aligned vertebrate coding sequences and then fixed. This allowed to capture many subtleties of sequence evolution that were not modeled by parametric models. However, the assumption of fixed substitution rates for all coding sequences was too stringent. Therefore, improved models have been developed, that combine the two approaches by taking empirical amino acid (Doron-Faigenboim and Pupko 2007) or codon (Kosiol et al. 2007) substitution rates and then adding parameters whose values can be estimated from the sequences under investigation.

The parameters used in these combined models were essentially still the same as in the purely parametric models, namely the codon frequencies π and the selection coefficient ω. The κ parameter had to be modified since these new models allowed substitutions between codons differing at more than one position so that combinations of two or three transitions or transversions can occur. But the basic assumptions made in 1994 about the necessary parameters in codon models have still not changed nor been challenged. Their choice is certainly reasonable and based on years of experience with single-nucleotide models of evolution. Using likelihood ratio tests and an Akaike information criterion (Akaike 1974), it was found that the ω parameter is very important for codon models, whereas κ varies little among genes and thus it is often not justified to reestimate this parameter (Kosiol et al. 2007; Doron-Faigenboim and Pupko 2007). However, to our knowledge, there is no study that conducted a systematic and unsupervised empirical search for the most suitable parameters of a codon model.

In this analysis we attempt to answer the question of the optimal choice of parameters for a codon model of evolution. With the tremendous growth of genomic data in the last few years, there are now enough sequences from many different species available to allow for an empirical study of the important aspects of codon evolution. As described in more detail in the "Methods" section below, we estimated many codon substitution rate matrices from alignments of Mammalian coding DNA and then performed principal component analysis (PCA) to find the most co-varying combinations of substitution rates.

Substitution rates that are expected to differ among data sets are modeled with parameters. Such a parameter not only allows for the estimation of certain rates from a specific data set, it also connects substitution rates that are expected to correlate strongly. The ω parameter, for example, is multiplied to all nonsynonymous rates, since it is assumed that the relative amount of nonsynonymous substitutions varies across different genes, but that this factor is approximately the same for all substitutions of this type. PCA (Pearson 1901) is a procedure to find groups of correlated variables. These sets of correlated variables are called principal components (PCs), and are defined such that no two PCs are correlated with each other. If PCA is applied to the substitution rates of a codon model, the resulting PCs describe sets of rates that co-vary the most and thus correspond to the optimal parameters for a substitution model.

The PCA is based on the covariance matrix C of the variables, which is constructed from many data points. The PCs are the eigenvectors of C and the corresponding eigenvalues indicate how much of the data's inherent variance can be explained by the respective PC. Any of the original data points can be reconstructed as a linear combination of the PCs. By using only a small number of PCs (those with the highest eigenvalues) reasonable approximations of all data points can be found. Thus, the PCA greatly reduces the number of variables needed to describe the distribution of the input data. This technique has been used in molecular sequence analysis before, e.g. Wang et al. (2008) performed PCA to identify recurrent patterns of amino acid frequencies across sites in alignments.

Methods

Alignments

The sequence data were taken from groups of orthologs from the OMA project (Dessimoz et al. 2005). It is not necessary to use orthologs for this task, but since the search for homologs is time-consuming, it is convenient to use the precomputed orthologs as available from OMA (Schneider et al. 2007). The Mammalian data set from 9 June 2009 was used, containing 62,156 groups covering 33 species. Since the alignments for this task should be of good quality and also contain enough data to estimate the many parameters, some filtering on the OMA groups was

performed. Coding sequences with more than one percent unknown bases were discarded. If in a group of orthologs more than 20 sequences were available, only the 20 longest were kept, while groups with less than 6 sequences were excluded. The remaining protein sequences of each group were aligned using the *Mafft* multiple sequence alignment (MSA) program (Katoh et al. 2005) and the corresponding coding DNA was then mapped to the aligned proteins. Of the resulting MSAs of DNA, all positions with a gap or where one of the codons contained unknown bases were removed. If after this treatment less than 1000 bases (333 codons) were left, the alignment was excluded from further analysis. Furthermore, if two or more of the remaining sequences were identical, only one of them was kept. This procedure resulted in 3666 MSAs of at least 6 sequences and 333 aligned codons. The guide trees required for the matrix estimation were constructed with the least-squares distance tree method from the *Darwin* bioinformatics package (Gonnet et al. 2000) using pairwise CodonPAM distance estimates (Schneider et al. 2005).

Substitution Rates Estimation

Estimating a 61 × 61 substitution matrix from a single MSA is a difficult task, since many parameters have to be estimated from relatively few observable substitutions. In order to obtain reasonable estimates with variances that are not too excessive, the number of parameters was reduced by assuming the codon frequencies as constant. For a time-reversible model, this leaves 1830 parameters, the off-diagonals of the so-called exchangeability matrix. This is a symmetric matrix which together with the equilibrium frequencies of the characters defines the substitution rate matrix (see, for example, Yang 2006).

The codon frequencies as well as an initial substitution matrix were estimated first from all 3666 MSA simultaneously. The initial substitution rate matrix was then used as the starting point for the estimation of the individual matrices from each MSA separately. All matrix estimation was performed with the expectation maximization (EM) method implemented in the *XRate* program (Klosterman et al. 2006) using a custom modification of the *codon.eg* grammar file to keep the codon frequencies fixed.

In order to verify if *XRate* was able to reliably estimate substitution rates from a single MSA, the parameter estimates from the 3666 separate matrices were compared to the corresponding parameter estimates from the initial matrix. The initial matrix can be considered as a reliable estimate as it was created using all MSAs simultaneously. For each of the 1830 parameters, the means and standard deviations were computed from the 3666 single-MSA estimates. The difference between the mean of the single-MSA estimates and the corresponding value from the initial matrix gives an indication of how well the parameters can be estimated from the individual MSAs. Figure 1 shows the distribution of these differences divided by the standard deviations. The vast majority of differences are smaller than one standard deviation and only 15 are larger than two standard deviations. This indicates that the parameter estimates from the individual MSAs lie within what can be expected.

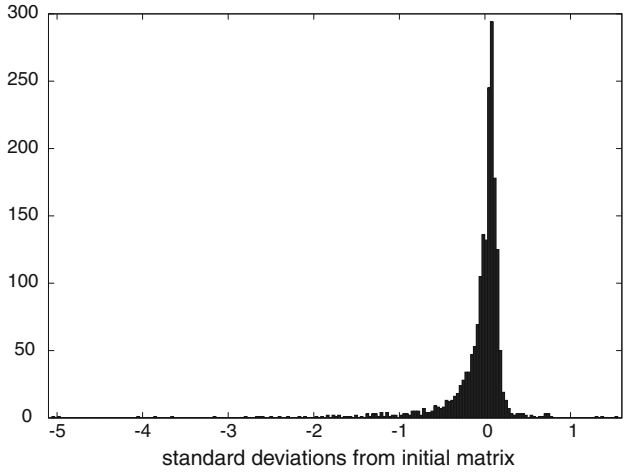

Fig. 1 Distribution of the differences between the means of the parameter estimates from single MSAs to the corresponding parameters from the joint estimation, divided by the standard deviation of the individual estimates

Principal Component Analysis

The PCA was performed with the *prcomp* function of the statistics software *R* (R Development Core Team 2009) on the 3666 estimated exchangeability matrices that were represented as vectors of length 1830 (the number of parameters). The resulting PCs (also vectors of length 1830) were then sorted by the corresponding eigenvalues and represented again as symmetric matrices, containing both positive and negative values.

Table 1 Selected eigenvalues from the PCA with their contribution to explain the variance and the cumulative contribution (sum of all contributions up to this number)

Number	Eigenvalue	Contribution (%)	Cumulative (%)
1	46.3	2.53	2.53
2	13.4	0.73	3.26
3	8.4	0.45	3.72
4	7.5	0.40	4.13
5	6.5	0.36	4.49
6	5.9	0.32	4.81
27	4.0	0.22	10.06
81	3.0	0.17	20.15
327	1.7	0.09	50.01

The *visualizeRates.pl* script that comes with the *XRate* program was used to display the PCs, after they were modified with a *ruby* script in order to allow for the display of negative values. Because the matrix is symmetric, it can be divided into two halves where the positive values are shown in the upper right half of the matrix and the negative values in the lower left half. The diagonal of the matrix can therefore be seen as a separation between two groups of parameters that are correlated with each other but anti-correlated with the parameters in the other half. The size of a circle is proportional to the corresponding parameter value. The larger a value, the higher is its influence on the PC.

Results and Discussion

Analysis of the Eigenvalues

The eigenvalues express how much the corresponding PCs contribute to explain the variance found in the input data set. The variance is made up of two parts: the true differences among the data points (here the different genes analyzed) and the noise in the data. It can be expected that the most important features are stronger than the noise and thus that the PCs with the highest eigenvalues are most likely to describe true features of the input data.

Table 1 shows a selection of the eigenvalues of the PCA performed in this study. The table displays the eigenvalue itself, its contribution to the explanation of the total variance as well as the cumulative variance explained by all PCs up to this one. The first PC has an eigenvalue of 46.2 and explains about 2.5% of the variance, which is more than three times that of the second PC. The second eigenvalue is 13.4 and explains only 0.7% of the variance. From there on, the distributions becomes much flatter with eigenvalues of 8.4, 7.5 and 6.5 for the next three PCs. It then takes 27 PCs to explain 10% of the variance and 327 for 50%. This indicates that at least the first two PCs explain real features, but from there on, the influence of the noise becomes stronger.

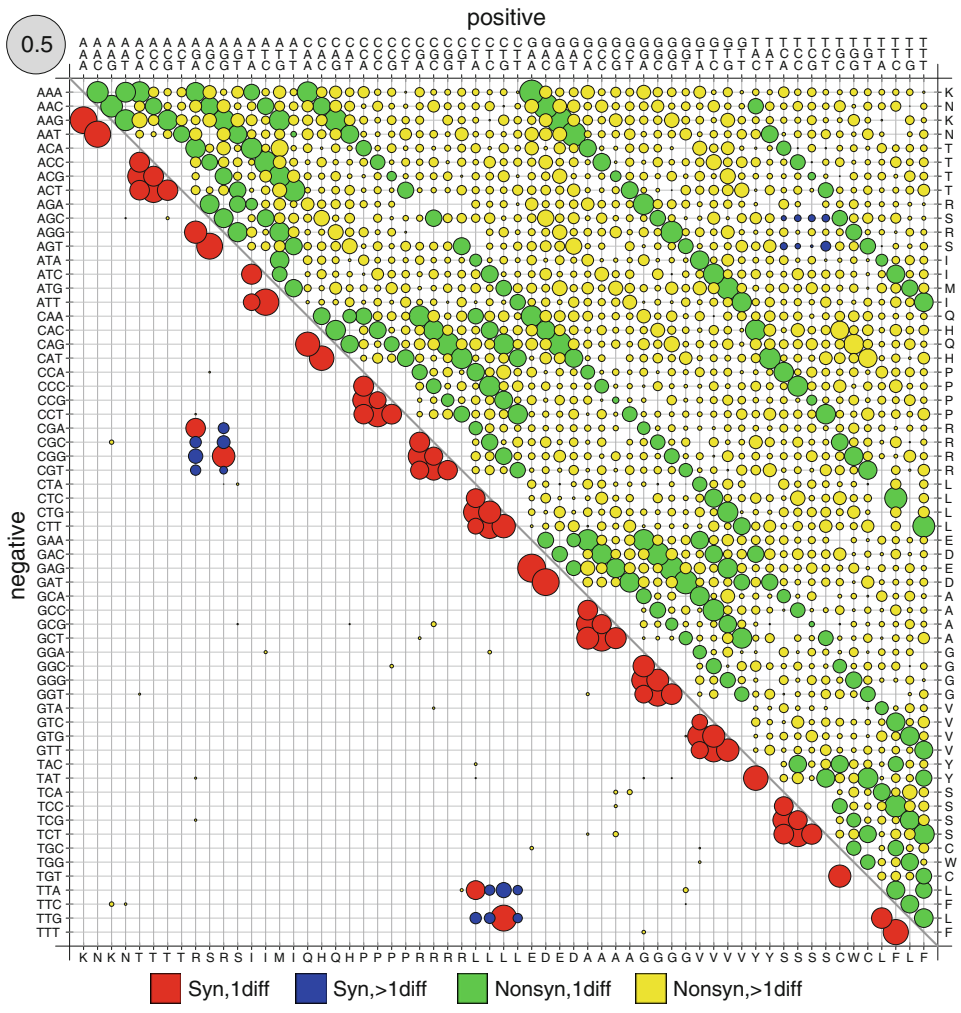

Fig. 2 First PC with positive values shown in the upper-right half and negative values in the lower left half of the matrix. The separation is clearly between synonymous (*red* and *blue*) and nonsynonymous (*yellow* and *green*) substitutions (Color figure online)

Analysis and Interpretation of the Principal Components

The first PC (the PC with the highest corresponding eigenvalue) is shown in Fig. 2, with the different substitution rates being coloured to indicate if they are synonymous (red and blue) or nonsynonymous (green and yellow). The figure shows an almost perfect separation between substitutions that are synonymous (in the lower-left half) and those that are nonsynonymous (in the upper-right half). There are a few nonsynonymous rates on the lower half, but the values of those parameters are very close to zero, indicating that their influence in the first PC is almost negligible. Interestingly, the only exception in the other direction are the synonymous substitutions within serine between AGC/AGT and TCN (N stands for any of the four bases) that are correlated with the nonsynonymous substitutions. The likely reason is that all of those substitutions require at least two nucleotide changes with the intermediate states encoding different amino acids. Thus, although the direct substitution from one codon to another would be synonymous, it probably often happens via intermediate nonsynonymous substitutions.

The second PC is shown in Fig. 3 with the same colour-coding as for the first figure. In this PC, the separation appears to be between substitutions where only one nucleotide differs (red and green) and substitutions with two or three nucleotide differences (yellow and blue). Unlike the close to perfect separation in the first PC, here there are several substitutions that do not follow the described pattern. However, the predominant separation still seems to be the number of nucleotide differences, with the lower left half clearly clustering the multi-difference substitutions independent of them being synonymous (blue) or nonsynonymous (yellow). Although the existence of positive instantaneous rates for multi-nucleotide substitutions are well documented (Averof et al. 2000; Whelan and Goldman 2004; Kosiol et al. 2007), it is a topic that is still not fully understood and subject to current debate (Anisimova and Kosiol 2009). Interestingly, the results

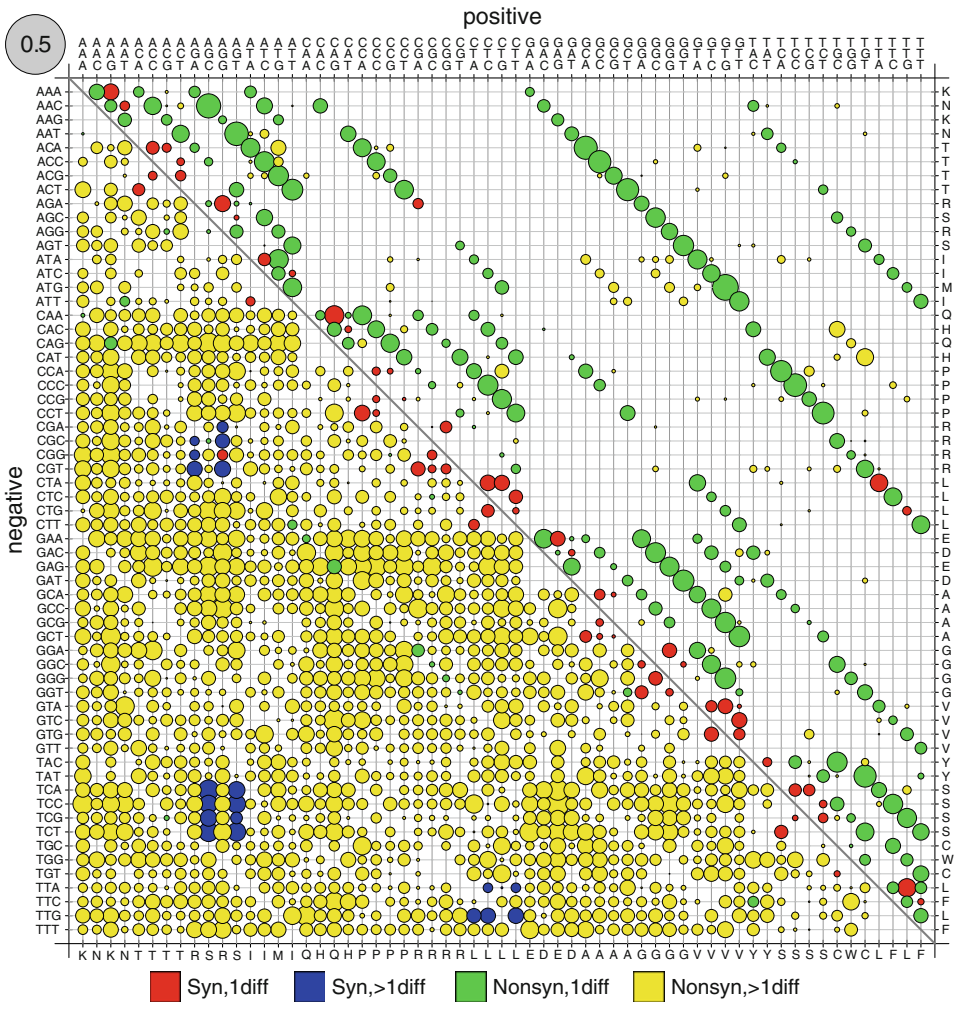

Fig. 3 Second PC, with *positive* values shown in the *upper right half* and *negative* values in the *lower left half* of the matrix. The separation is mostly between single-nucleotide substitutions (*red* and *green*) and substitutions involving two or three nucleotide changes (*yellow* and *blue*) (Color figure online)

from our study suggest that the extent of multi-nucleotide substitutions is one of the most important parameters for codon substitution models.

The four next PCs are shown in Fig. 4. No simple description could be found for any of them, but at least in the third and fourth PCs, a grouping of substitutions involving the same amino acids can be noticed and the separation occurs almost entirely among nonsynonymous substitutions. This could mean that these PCs are caused by some physico-chemical features of the amino acids. But in any case, the lack of an obvious description for these PCs implies more complex parameters, which so far have not

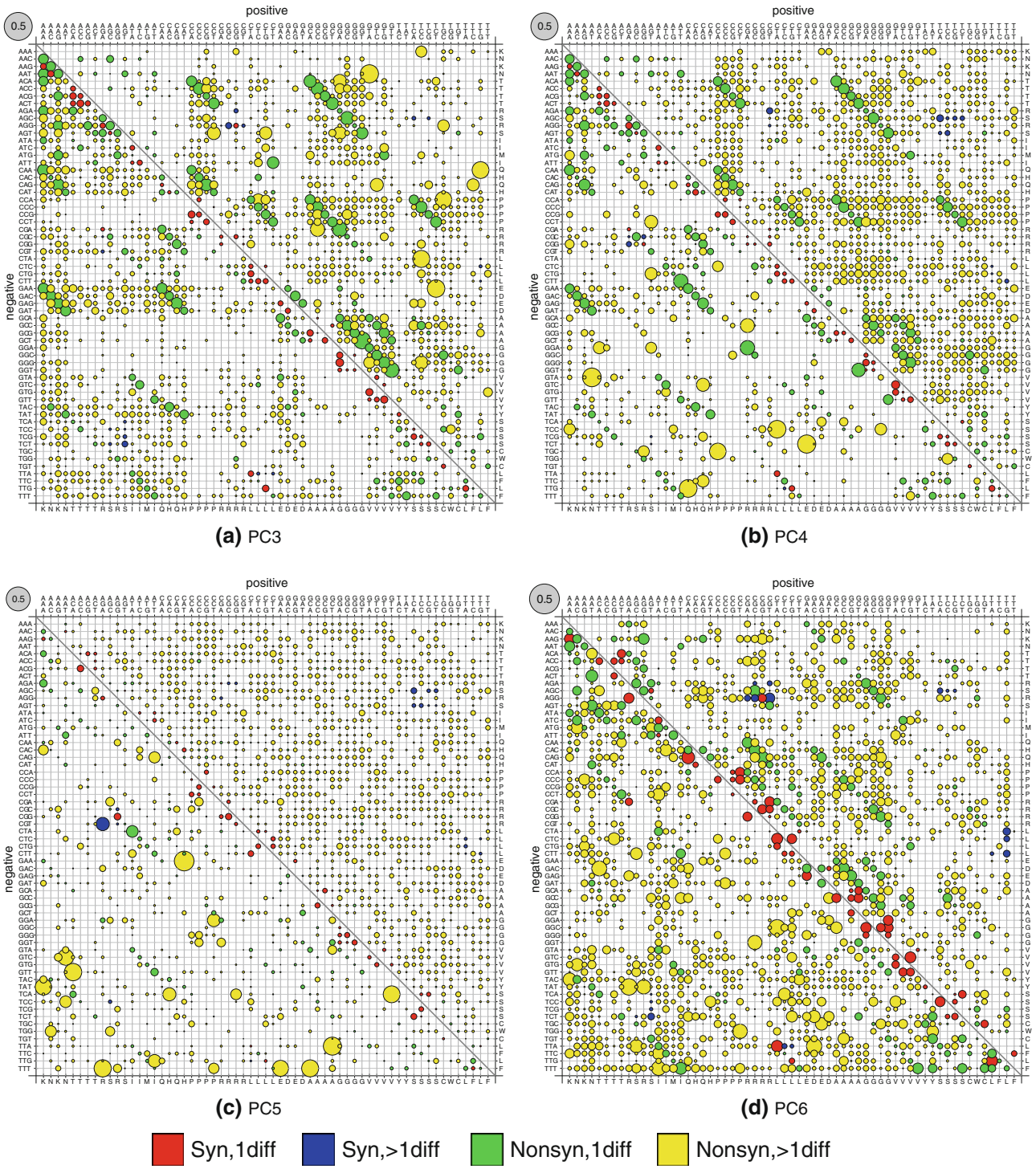

Fig. 4 Combined figure of the third (**a**), fourth (**b**), fifth (**c**) and sixth (**d**) PCs (Color figure online)

been considered in parametric models. Also, there is always the problem of noisy data. Since the basis for this analysis were substitution matrices estimated from single MSAs, considerable amounts of variance must be taken into account. It is difficult to distinguish between true parameters (as the first two PCs clearly were) and the influence of noise that could play a role for the higher PCs.

Conclusions

The PCA of codon substitution matrices estimated from 3666 MSA allowed for an empirical analysis of the important parameters of codon substitution models. That the first two PCs clearly identify previously described features of codon evolution, namely the ratio of nonsynonymous to synonymous substitution rates and the rate of multi-nucleotide substitutions, is a confirmation that substitution matrices estimated from single alignments are reliable enough for this type of analysis.

The almost perfect correspondence of the first PC to the selection coefficient ω clearly supports the importance of this parameter, as it is the parameter that varies the most among genes. The only notable exceptions are the intra-serine substitutions that require nonsynonymous intermediate changes and that are clustered with the nonsynonymous substitutions. This could indicate that those substitution rates might be better modeled as nonsynonymous rates in parametric models.

Interestingly, the separation of transitions and transversions is not reflected in any of the first six PCs, most likely because there is not much variation among coding sequences in terms of the κ parameter. This would imply that this parameter might not be a significant contribution to a codon substitution model and could also be fixed to an average value in order to reduce the number of parameters. Both Kosiol et al. (2007) and Doron-Faigenboim and Pupko (2007) came to similar conclusions concerning this parameter.

It might be surprising that the amount of multi-nucleotide substitutions is the second most important factor. But since this PC is well defined, it appears to be a real signal. The mechanisms of multi-nucleotide substitutions are still not fully understood, but this at least shows their importance. It is also noteworthy that since PCA finds the factors that vary the most within the data set, there could be an evolutionary mechanism that influences the amount of multi-nucleotide substitutions that can get fixed in the evolution of a coding sequence.

Overall, and even though the third and subsequent PCs did not show a clear pattern, this study can contribute to the understanding of the factors involved in DNA evolution. This could lead to better codon models, in particular semi-parametric models with better parameters, such as treating intra-serine substitution as nonsynonymous, fixing the value of the transition/transversion rate ratio or using parameters to model multi-nucleotide substitutions. But it could also lead to a new generation of empirical models where some of the PCs found here or factors found by similar methods are directly integrated into the model.

Acknowledgements We would like to thank Gina Cannarozzi and Gaston Gonnet for helpful discussions as well as the two reviewers for their valuable comments. Adrian Schneider is supported by a grant from the Swiss National Science Foundation (SNF).

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automatic Control 119:716–723

Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26(2):255

Averof M, Rokas A, Wolfe KH, Sharp PM (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science 287(5456):1283

Dessimoz C, Cannarozzi G, Gil M, Margadant D, Roth A, Schneider A, Gonnet G (2005) OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In: McLysath A, Huson DH (eds) RECOMB 2005 Workshop on Comparative Genomics. Lecture Notes in Bioinformatics, volume 3678. Springer-Verlag, Berlin. pp 61–72

Doron-Faigenboim A, Pupko T (2007) A combined empirical and mechanistic codon model. Mol Biol Evol 24(2):388–397

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11(5):725–736

Gonnet GH, Hallett MT, Korostensky C, Bernardin L (2000) Darwin v. 2.0: an interpreted computer language for the biosciences. Bioinformatics 16(2):101–103

Katoh K, Kuma K, Toh H, Miyata T (2005) Mafft version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33(2):511–518

Klosterman P, Uzilov A, Bendaña Y, Bradley R, Chao S, Kosiol C, Goldman N, Holmes I (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics 7:428

Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24(7):1464–1479

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11(5):715–724

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148(3):929–936

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2(6):559–572

R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Schneider A, Cannarozzi GM, Gonnet GH (2005) Empirical codon substitution matrix. BMC Bioinform 6:134

Schneider A, Dessimoz C, Gonnet GH (2007) OMA Browser—exploring orthologous relations across 352 complete genomes. Bioinformatics 23(16):2180–2182

Wang H-C, Li K, Susko E, Roger AJ (2008) A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol 8:331

Whelan S, Goldman N (2004) Estimating the frequency of events that cause multiple-nucleotide changes. Genetics 167(4):2027–2043

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15(5):568–573

Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19(6):908–917